Identifying Potential Positive Deviants (PDs) Across Rice Producing Areas in Indonesia: An Application of Big Data Analytics and Approaches

Pulse Lab Jakarta Follow
Jun 22 · 8 min read

Authors: Basma Albanna, Data Science Fellow, Pulse Lab Jakarta — Doctoral Researcher, University of Manchester; **Dharani Dhar Burra**, Data Scientist, Pulse Lab Jakarta; and **Michael J. Dyer**, Geospatial Information Systems Officer, Pulse Lab Jakarta.

Approaches to data collection for development programming have commonly relied on official statistical data in combination with surveys to plan, implement and monitor progress and impact for interventions. Today, the emergence of big data has resulted in a paradigm shift, with increasing use of non traditional data to promote more effective and responsive interventions across various domains. Contributing to the global ***Data Powered Positive Deviance initiative****, Pulse Lab Jakarta conducted data analytics research by merging traditional statistical data with Earth Observation big data to identify potential rice producing villages across Indonesia that might be faring better than others, referred to as positive deviants (PDs). Our team recently wrapped up the pilot study and this post is intended to capture the process that they* *undertook. We are also pleased to share the* ***technical report****, detailing the preliminary results, key learnings, along with some actionable recommendations for future work in the agriculture domain.*

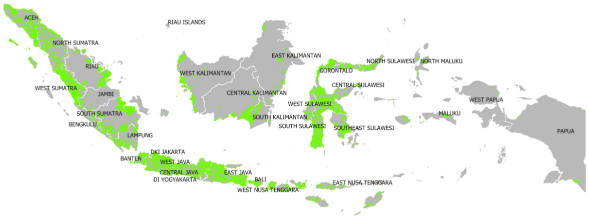
Positive Deviance Approach

The Positive Deviance approach is focused on identification and scaling of strategies undertaken by positive deviants (PDs), which refer to individuals or communities that use uncommon practices that enable them to achieve better outcomes than their peers, despite having similar conditions and resources. This bottom-up approach for development programming has been adopted with relative success across a range of sectors, such as public health and agriculture. Although the approach has had success, the scaling of successes achieved across diverse geographies and large populations has presented numerous challenges, part of which can be attributed to the conventional qualitative and quantitative approaches employed, namely interviews and surveys. Recent developments in technologies and the availability of big data have also presented new possibilities for the PD approach, whereby big data can be harnessed to fill those spatio-temporal information gaps.

Developing A Statistically Robust Method

Statistically rigorous, time-tested methods and processes underline the generation of official statistical data. As such, it's reasonable that trialling a new approach such as Positive Deviance that combines official statistical data with big data would undergo similar scrutiny. Our pilot project, was premised on this understanding, wherein the focus was to develop a statistically robust method that can re-analyse exemplar official statistical data (agriculture census and village potential survey), in combination with open access Earth Observation (EO) big data, using a PD-based framework.

For this research, re-analysis entailed leveraging the aforementioned data sets to identify communities of rice farming villages, and then within those communities, to further identify individual villages with relatively high agriculture productivity (high performers). This was then followed by getting a sense of successful practices that might have been responsible for high agriculture productivity, as identified using official statistical data. Lastly, we sought to identify opportunities that may help to eventually scale successful practices across remaining individual villages that have common bioclimatic conditions.

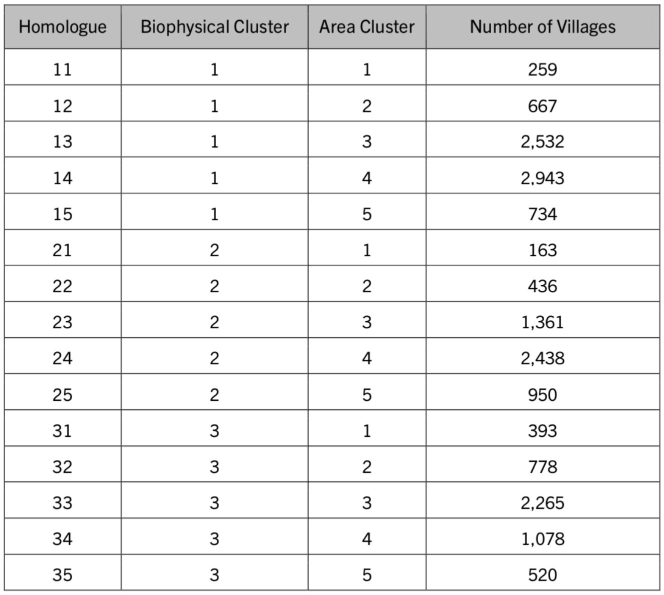


Rice growing areas across Indonesia (2014), including both wetland and dryland production.

Whilst these components are discussed at length in the **technical report**, we thought it'd be useful to highlight some of the challenges we encountered and practical solutions we engineered throughout the development of our method:

Construction of Homologous Environments

In a PD-based approach, a homologous environment (HE) refers to a group consisting of entities with similar characteristics. Given the diverse bio-geographic conditions across the Indonesian archipelago, determining such environments is challenging due to extremely diverse agricultural production systems. In addition, agricultural productivity is often contingent on conditions that may, and in other cases, may not be controlled by individual farmers. For the purpose of our research, we defined HEs as those that had experienced similar bioclimatic conditions (temperature and precipitation) during the cropping season being examined. With temperature and precipitation exemplifying conditions that cannot be directly controlled by farmers, we hypothesised that: *if homologous environments were constructed based on these uncontrollable conditions, then the residual should indicate conditions (and practices) that are controllable.* We were able to characterise agricultural production systems across Indonesia into 15 distinct HEs. Using machine learning methods, this process integrated bioclimatic data of the target cropping season, derived from open access EO datasets.



Homologue	Biophysical Cluster	Area Cluster	Number of Villages
11	1	1	259
12	1	2	667
13	1	3	2,532
14	1	4	2,943
15	1	5	734
21	2	1	163
22	2	2	436
23	2	3	1,361
24	2	4	2,438
25	2	5	950
31	3	1	393
32	3	2	778
33	3	3	2,265
34	3	4	1,078
35	3	5	520

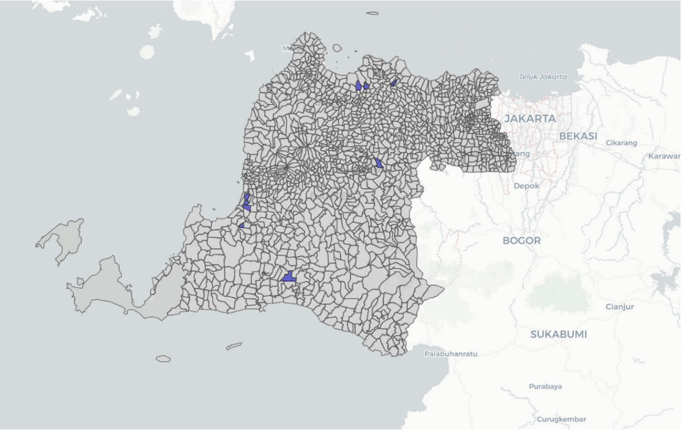
Among the 15 distinct HE identified, the table shows the number of villages within each.

Selection and Measurement of Performance

This step was particularly challenging for us, because we realised that the selection of a performance measure is not only subjective and context dependent, but in our case was also dependent on the available data. Furthermore, performance measures are latent, temporally sensitive, with multidimensional constructs. Essentially, these performance measures cannot be measured directly, and thus may need to be derived based on various measures. For example, high agricultural productivity at the expense of intense agro-chemical use may not always be considered high performance if we were to include environmental sustainability as an additional parameter. As the census is not designed to capture information on yields, this also presented a challenge in terms of the data we had at hand. To circumvent this issue, we selected Enhanced Vegetation Index (EVI), a commonly used measure for productivity that was derived from open EO data, as a proxy indicator.

Identification of Outliers

Once the performance measure was established, our next step was to figure out how to determine the high performers based on this measure. There is a general notion that as more data becomes available, the easier it is to identify outliers. The challenge however is that big data also comes with large amounts of irrelevant information (called noise), and therefore it is necessary to use methods that can filter noise from signals with a certain degree of confidence — this in particular was central to this pilot study and will be pertinent for other Data Powered Positive Deviance projects. We employed distributional cut-offs, partial least squares regression, in combination with a variety of outlier detection methods, to not only identify high performers/PDs, but also structural variables and conditions (derived from official statistical data) associated with the high performers within each HE. It's interesting to note that we found several high performing villages across a number of HEs with relatively younger farmers, lower incidences of flooding events and followed a combination of plantation farming along with rice farming. Also of interest, these villages were dependent on rain-fed systems. However, to revert to the point made earlier on performance being subjective, these observations would be more valid and reliable, if high productivity itself was considered as the performance measure. Due to these inherent biases of performance measures, we chose to use the term "outliers" instead of "high performers" throughout the report.



Outlier identification is an integral part of the proposed method. We experimented with multiple approaches for outlier identification, and only those outliers that were found across more than one outlier method were termed as "True Outliers". Mapped (in dark blue) here are "True Outlier" villages of Cipeundeuy, Sobang, Kubangkampil, Sukaresmi Pagelaran, Rangkasbitung, Tanara, Cigelam, Pontang identified in the province of Banteng.

Outlier Validation

Now that we've identified these outliers, how do we validate whether what we identified was random noise or actual signals? This was an important question we needed to answer in order to effectively communicate the findings and potential utility to government stakeholders, to then promote scaling up PD based approaches for development interventions. We implemented three different validation approaches, all of which assessed whether the identified outliers are indeed outliers, or artefacts of the method used. Detailed in the technical report, two of the three validation approaches relied on reviewing literature that investigate the relationship between the structural variables significantly associated with outliers and high agriculture productivity. To complement this literature review, we also used the Google time scale tool, and analysed historical satellite imagery by searching for evidence for these successful practices (for instance searching for evidence of plantation farming), on a subset of villages across all HEs. While these validation approaches were dependent on finding evidence for successful practices either through literature review or through satellite imagery, we also developed and implemented another validation test, which relied only on time series EO data. These three validation approaches revealed that the outliers identified are not artefacts of the method (not

Pulse Lab Jakarta is grateful for the generous support from the Government of Australia

Big Data Agriculture Big Data And Analytics Positive Deviant Rice

noise); the outlier identification method instead did pick up actual signals and identified “true” outliers (potential positive deviants).

Moving Forward

From the start, our team set out to build a robust, statistically rigorous method that uses official statistical data in combination with EO data to identify high performing rice farming villages. In order to progress further, we acknowledged that obtaining promising results during the validation steps would be necessary. Primary data collection activities are expensive and laborious, and are thus associated with costs proportional to the sample size. For this reason, it is important to weigh the costs and the associated benefits before choosing to perform additional ground-based validations. Now that the method is able to identify, and geospatially locate outliers with relatively high confidence, the next step would be to physically visit a subset of outlier and non-outlier villages, and perform a final round of ground-based validation. However in this instance since the official statistical data utilised for this preliminary exercise was more than five years old, it would be more prudent (and subject to buy-in from government stakeholders) to repeat this analysis with the upcoming agriculture census and Village Potential Statistics survey (PODES). Once this analysis is repeated with more recent data, it should be followed up by going the additional step of conducting ground truthing through a PD inquiry (ground survey and ethnographic methods targeting the true outliers to understand their underlying behaviours)[1]. That would then set the stage for the scaling of these practices across their respective HEs.

[1] Pascale, Sternin, & Sternin. (2010) The Power of Positive Deviance: How Unlikely Innovators Solve the World’s Toughest Problems. Harvard Business Press. Print.

Data Powered Positive Deviance is a global initiative collaboratively created by GIZ Data Lab, Pulse Lab Jakarta, UNDP Accelerator Labs Network, and the University of Manchester Centre for Digital Development. If you’re interested in knowing more about our pilot research aimed at identifying potential positive deviants (PDs) across rice producing areas in Indonesia, get in touch with our team plj@un.or.id

Editors: Dwayne Carruthers, Communication Manager; and **Utami Diah Kusumawati**, Communication Assistant, Pulse Lab Jakarta.

· · ·